混合蛙跳算法在文本分类特征选择优化中的应用*

路永和　陈景煌

(中山大学资讯管理学院　广州 510006)

摘要:【目的】由于文本数据存在许多与分类不相关的冗余词项，引入混合蛙跳算法进行特征选择优化，提高分类准确率。【方法】分别使用 CHI 和 IG 预选出不同维度的特征集合，再引入改进后的混合蛙跳算法对预选特征集合进行二次优选，每只青蛙的位置代表一种特征选择规则，将分类准确率作为算法的适应度函数。SVM 和 KNN 分类器用于实验中分类准确率的计算。【结果】引入改进后的蛙跳算法比 CHI 和 IG 能得到更好的分类效果，最大提升幅度达到 12%。【局限】在少部分特征维度下出现过拟合现象。【结论】采用特征词预选和改进后的蛙跳算法相结合的特征选择优化方法可以有效排除部分噪声特征项的干扰，从而提高文本分类准确率。

关键词: 特征选择　文本分类　混合蛙跳算法

分类号: TP391

1　引　言

在文本信息处理领域，文本分类作为信息挖掘、自然语言处理、信息检索等技术的重要基础[1]，得到了许多学者的关注和研究。文本分类技术已经从传统的人工分类发展到基于机器学习的自动分类[2]，文本分类在质量和效率两方面都得到较大提高。而文本数据往往具有高维、稀疏、多标号等特点，这些在一定程度上影响了文本分类效果，因而文本特征选择优化成为学界的研究热点。在向量空间模型(Vector Space Model, VSM)中，原始特征集合中的每个特征项对分类学习不一定都是必要的，有些噪声特征项不仅增加了特征集合的维度，而且会影响文本分类的整体效果。因此需要对特征集合进行降维处理。

本文使用在文本领域还未得到较多应用的混合蛙跳算法(Shuffled Frog Leaping Algorithm, SFLA)，对其进行编码规则、个体进化方式等方面的改进，并将其应用在文本特征选择优化中，最后通过实验证明这种方法的有效性。

2　相关研究

2.1　传统的文本特征选择方法

文本分类的过程主要包括：文本预处理和分词、文本表示、特征选择、权重计算、使用分类器分类。其中，文本表示主要是采用 VSM 表示[1]，而文本经预处理后得到的特征集合的维数非常高，特征分布稀疏，因此每个文本都被表示成一个高维向量。而高维向量对分类器造成很大的计算负担，因此文本特征选择在文本分类中非常重要，经过特征选择后得到具有文本代表性的特征词集合，从而降低每个文本向量的空间维数，提高分类效率和准确率。目前学界使用的特征选择方法主要有文档频率(Document Frequency, DF)、卡方检验(CHI)、信息增益(IG)、互信息(Mutual Information, MI)等。有相关试验证明，CHI 分类效果好但是计算开销较高[3]；在英文文本集的分类中，CHI 与 IG 效果最佳，DF 基本与前两者相当，而 MI 则相对较

通讯作者：路永和，ORCID：0000-0002-7758-9365，E-mail：luyonghe@mail.sysu.edu.cn。

*本文系国家自然科学基金项目“面向文本分类的多学科协同建模理论与实验研究”(项目编号：71373291)和广东省科技计划项目“面向主题的中文语料库构建方法与技术”(项目编号: 2015A030401037)的研究成果之一。

差[4]; 在中文文本集的分类中, CHI 的效果最佳, 其次为 IG, 而 MI 相对较差[5], DF 的效果居中[3]。

但是 CHI、IG 等传统的特征选择方法是通过某一数学模型从原始特征集合中筛选出具有较好的区分能力和文本代表性的特征集合, 并没有从文本的角度考虑特征词之间的相互影响以及冗余词项对文本分类效果的整体影响。因此, 基于传统特征选择方法, 通过引入改进后的混合蛙跳算法, 利用该算法较强的寻优能力, 对预选的特征集合进行二次优化, 从而得到特征维度相对较低的高精度特征集合, 并且改进了最终的分类结果。

2.2 结合群体智能算法的特征选择优化

近年来, 不断有学者将群体智能算法应用到文本特征选择领域中, 并且效果明显。总体方向大致可以分为两个:

(1) 直接使用群体智能算法进行文本特征选择, 不再使用传统文本特征选择方法, 这个方向的研究成果主要有: Tabakhi 等[6]提出 UFSACO 方法, 即将蚁群算法(ACO)引入到无监督的特征选择方法中, 考虑到特征之间的相关性, 从而提出特征集合中的冗余词项, 实现降维效果, 并通过实验说明该方法比传统特征选择方法能得到更好的分类效果。刘亚南[7]将基于遗传算法(GA)的文本特征选择方法运用到动态获取 K 值的 KNN 分类算法中。刘逵[8]构建基于野草算法的文本特征选择模型, 该模型可以给予权重值较低的词条进行特征选择的机会, 同时保证权重值高的特征词选择优势, 从而更全面地提高文本特征选择的全面性和准确率。

(2) 将群体智能算法结合传统文本特征选择方法, 即先使用传统特征选择方法得到预选特征集合, 再引入群体智能算法进行精选, 最后得到高精度的特征集合, 从而提高文本分类效果, 主要有以下研究成果: Uguz[9]在使用传统特征选择方法 IG 的基础上, 分别引入遗传算法和主成分分析法(PCA)进行二次特征选择和抽取, 剔除与分类无关的特征词项, 实现降维, 并且取得不错的分类效果。Javed 等[10]通过使用传统特征选择方法 BNS 和 IG 进行特征词预选, 然后结合 Markov Blanket Filter(MBF)算法对预选特征词进行二次筛选, 从而实现降维并改进了文本分类效果。Lu 等[11]使用 CHI 进行特征词预选, 然后分别使用所提出的 6 种改进的粒子群优化算法(PSO)对预选特征集合进行精选, 最后通过实验表明异步改进的 PSO 算法具有最佳的文本分类效果。

本文将 SFLA 结合传统文本特征选择方法, 先进行特征词预选, 再引入改进后的二进制 SFLA 进行特征词精选, 从而得到高精度的特征集合, 并最终改进文本分类效果。

2.3 混合蛙跳算法

混合蛙跳算法是由 Eusuff 等[12]提出的一种协同搜索群智能算法,该算法同时结合了模因算法(Memetic Algorithm, MA)和粒子群优化算法, 既有模因算法的遗传特性, 又有粒子群算法的社会信息共享的特点。算法流程简单合理, 参数较少, 并且收敛速度快、全局寻优能力强。

SFLA 最初受青蛙觅食的生物现象启发而被提出。由 N 只青蛙组成的蛙群 P 在一个受限的 S 维度空间中寻找有限且最优的食物源。每只青蛙 i 的位置用 $X_i=\{x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{iS}\}$, 其中 S 表示青蛙所在空间的维度, X_i 在对于解决优化问题时则表示一个可行解向量, 并计算每只青蛙当前位置的优劣程度, 即适应度 $F(X_i)$。然后按照其适应度 $F(X_i)$大小降序排列, 并记录当前种群的全局最优位置 X_g。再将整个蛙群分成 n 个族群, 每个族群包括 m 只青蛙。分组规则为: 第 1 只青蛙分入第 1 个族群, 第 2 只青蛙分入第 2 个族群, 第 m 只青蛙分入第 m 个族群, 第 m+1 只青蛙分入第 1 个族群, 以此类推, 并记录每个族群的局部最优解 X_b 和最差解 X_w。接下来每个族群进行组内进化, 进化的方式[12]为:

$$D=rand()\cdot(X_b-X_w) \tag{1}$$

$$X'_w=X_w+D,\quad -D_{\max}\leqslant D\leqslant D_{\max} \tag{2}$$

其中, $rand()$为 0 到 1 之间的随机数; D 是指青蛙每次跳跃的步长距离, X'_w 是指跳跃后青蛙所处的位置。

根据公式(1)和公式(2)计算得出 X'_w。如果 X'_w 适应度 $F(X'_w)$ 优于 X_w 的适应度 $F(X_w)$, 则用 X'_w 代替 X_w, 继续下一次的组内进化; 否则用 X_g 代替公式(1)中的 X_b, 根据公式(1)和公式(2)计算得出 X'_w。如果 X'_w 的适应度 $F(X'_w)$ 优于 X_w 的适应度 $F(X_w)$, 则用 X'_w 代替 X_w, 进入下一次的组内进化; 否则随机生成一个 X'_w, 并用其代替 X_w, 进入下一次组内进化。当每个族群的组

内进化次数都达到最大次数 L 时，将所有族群的青蛙重新混合在一起，重新按照各自的适应度 $F(X_i)$降序排列，更新当前最优解 X_g，并以此种群为基础，继续构造下一代新种群，直到达到最大总迭代次数 T 或者满足算法结束条件[13]。

目前 SFLA 已经被应用到水资源网络优化[12]、桥面修复[14]、含风电场电力系统的动态优化潮流计算[15]、分布式风电源(DWG)规划模型[16]、语音识别[17]等领域中。

但是在所查找的文献中，SFLA 被应用于文本信息处理领域的相关研究较少。其中，许方[18]改进了传统的 SFLA，并将其分别与 K-means 和 FCM 结合，应用到文本聚类领域中，并且提高了 Web 文本聚类的精度。同样在文本聚类方面，尉建兴等[19]将 SFLA 与 K-means 算法结合，提高了聚类的性能。在文本分类方面，Sun 等[20]则以 SFLA 直接作为分类算法，以 LDA 作为特征选择方法，提高了 Web 文本分类的准确率。截至目前，SFLA 在文本信息处理领域中的应用比较少。本文尝试对 SFLA 进行改进，将其与传统特征选择方法结合，并通过实验验证其有效性与可行性。

3 基于混合蛙跳算法的文本特征选择优化

3.1 算法改进

(1) 编码规则

由于文本特征选择优化问题本质上是组合优化问题，所以 SFLA 将进行二进制编码规则改进，即每一只青蛙对应的位置代表一种特征选择规则，一只青蛙的每一维对应一个特征项，而每一个特征项对应着两种结果：被选中与不被选中，每个特征项被选中则取 1，不被选中则取 0。所以，每个解向量(青蛙的位置)可以表示为：

$$X_i=\{x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{iS}\},\quad x_{ij}\in\{0,1\} \tag{3}$$

其中，X_i 表示第 i 个解向量，x_{ij} 表示第 i 个解向量的第 j 个分量，并且只可以取 0 或者 1。若 $x_{ij}=1$，说明第 i 个解向量中的第 j 个特征项被选中；若 $x_{ij}=0$，说明第 i 个解向量中的第 j 个特征项未被选中。

(2) 个体进化方式的改进

由于本文使用的 SFLA 是二进制编码，标准 SFLA 的个体进化方式(即公式(1)和公式(2))不再适用，因此对 SFLA 的个体进化方式做如下改进，使其能够更适用于文本特征选择的优化，具体改进流程如图 1 所示。

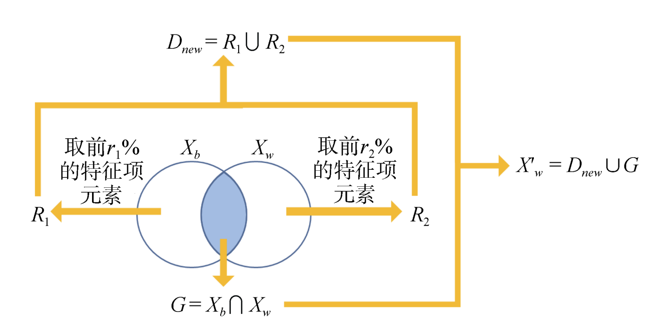


图 1 SFLA 的个体进化方式改进流程图

首先求出某个族群里的最优解 X_b 和最差解 X_w 都选中的特征项集合 G(即对于第 j 个分量(特征项)，X_b 与 X_w 同时取 1 的所有分量的集合)，将 X_b 与 X_w 看作是集合，则 G 是 X_b 与 X_w 的交集：

$$G=X_b\cap X_w \tag{4}$$

然后求每只青蛙跳跃时的步长 D_{new}，计算公式如下：

$$R_1=r_1\odot(X_b-X_w) \tag{5}$$

$$R_2=r_2\odot(X_w-X_b) \tag{6}$$

$$D_{new}=R_1\cup R_2 \tag{7}$$

其中，(X_b-X_w)与(X_w-X_b)表示集合的差运算。r_1 与 r_2 是 0 到 100 的随机整数，$r_1\odot(X_b-X_w)$ 表示从(X_b-X_w)这个集合里取前百分之 r_1 的特征项元素，构成集合 R_1，$r_2\odot(X_w-X_b)$ 表示从(X_w-X_b)这个集合里取前百分之 r_2 的特征项元素，构成集合 R_2；再取二者并集得到集合 D_{new}，即为每只青蛙跳跃的步长。如：当 $r_1=20$，$r_2=40$，(X_b-X_w)集合中有 100 个元素，(X_w-X_b)集合中有 200 个元素，则从(X_b-X_w)集合中取前 100×20%=20 个特征项，从(X_w-X_b)集合中取前 200×40%=80 个特征项，这 20+80=100 个特征项组成了集合 D_{new}，即某只青蛙某次跳跃时的步长。最后组内某只青蛙某次跳跃后的位置更新为：

$$X'_W=G\cup D_{\rm new} \tag{8}$$

这里对 SFLA 的个体进化方式改进是基于以下理由：首先求最优解 X_b 与最差解 X_w 之间的交集 G，即保留二者之间的“共同特征项”，从而新产生的个体在“继承”其二者的共同特征项的基础上继续进化，寻找到更优位置。然后计算青蛙跳跃时的步长时，分别从 X_b 与 X_w 各自“特有”的特征项元素中选取前若干个特征项来组成集合 D_{new}。这样的做法是让新产生的个体随

机“继承”若干比例的 X_b 与 X_w “特有”的特征项，从而让新个体产生某个方向的进化；另外，由于候选特征集合是经过CHI或者IG筛选得到的，集合中的特征项都是按照CHI得分或者IG得分从高到低排序的，得分越高则越有代表性，所以选取的是排位靠前的若干个特征项。

(3) 最大移动步长 D_{max} 的改进

以上对标准 SFLA 的个体进化方式中步长的计算进行了改进，使其适用于解决特征选择优化问题，所以对最大移动步长 D_{max_new} 也需要进行重新定义。

首先定义一个新变量：差异度(diff)，是指新产生个体 X'_w 与原来 X_w 之间在对应维数的解分量上存在多大比例不同；则 D_{max_new} 指允许新产生个体 X'_w 与 X_w 之间的最大差异度。比如：$X'_w=\{1,0,1,1,0,1\}$，$X_w=\{0,1,1,1,0,0\}$，X'_w 与 X_w 分别在第 1、2、6 维的解分量上不同，则二者的差异度 diff=(3/6)×100%=50%，所以二者存在 50%的差异。

引入差异度 diff 这个变量是为了计算二进制编码规则下的青蛙个体之间的差异比例，相当于标准 SFLA的步长；但由于对二进制SFLA下的步长的计算公式进行了改进，步长不再表示新个体与原来个体之间的差异程度。因此改进后的蛙跳算法的最大移动步长 D_{max_new} 是指允许新产生的个体 X'_w 与原来的 X_w 之间的最大差异度。

3.2 相关参数设置

本文采用的改进后的二进制 SFLA 算法共需要设置 5 个参数：蛙群规模 N、族群数量 n、最大移动步长 D_{max}、族群内进化次数 L、总迭代次数 T。参数的设置对算法的运行效果有较高的影响程度。

SFLA 的蛙群规模是指种群中所有青蛙的数量 N，对于组合优化问题则是指初始生成的解向量个数。一般情况下，N值与问题的复杂度相关，但由于本实验在计算青蛙的适应度的时间开销较大，因此将青蛙总数量设置为 20。SFLA 的族群 n 要根据划分后每个族群内青蛙的数量 m 的大小来设置，本文将族群数量 n 设置为5，则族群内青蛙数量为4。改进后的二进制SFLA的最大移动步长 D_{max} 是指允许新产生的个体与原来个体在对应解向量上的最大差异程度，在作用上与标准 SFLA 中的 D_{max} 是相似的，都是为了控制算法进行全局搜索的能力。实验将 D_{max} 设置为 45，即新产生个体与原个体在对应解向量上的差异度不得超过 45%。参数 L 决定着族群内青蛙的进化次数；总迭代次数 T 主要与问题的复杂度相关，问题复杂度越高，T 也应设置得越大，找到最优解的概率才会增大。但由于实验计算青蛙适应度的时间开销较大，故将族群内迭代次数 L 设置为 10，将总迭代次数 T 设置为 10。

3.3 适应度函数

群体智能算法的适应度函数用来计算个体的适应度，一般是由算法的优化目标来决定。本文引入 SFLA 对特征选择进行优化，主要目标是降低文本特征集合的维度以及提高文本分类的准确率。因此，将文本分类准确率作为衡量每只青蛙所处位置的优劣，使青蛙向分类准确率更高的位置“跳跃”，即：

$$Fitness()=\frac{\text{分类正确的测试文本数}}{\text{测试文本集中的文本总数}} \tag{9}$$

3.4 算法设计

基于改进后的 SFLA 的文本特征选择优化算法流程如下：

输入：训练文本集 TR，测试文本集 A，通过 CHI 或 IG 要得到的预选特征词数量即特征空间维度 S，初始化的青蛙数量 N，族群数量 n，最大移动步长 D_{max}，族群内最大进化次数 L，总迭代次数 T。

输出：经过 SFLA 二次优选的特征集合。

(1) 使用分词软件对训练文本集 TR 进行分词处理，然后分别使用 CHI 和 IG 进行文本特征预选择，得到候选特征集合；

(2) 使用随机函数从{0, 1}为蛙群中每只青蛙的位置的每一维度选定一个值，对应维度的值为 1 则表示选择该特征词，对应维度的值为 0 则表示不选择该特征词，以此作为每只青蛙的位置初始值；

(3) 计算每只青蛙所处位置的适应度，即分类准确率。将每只青蛙的位置的各个维度上值为 1 的特征词作为测试文本集 A 的特征表示，构造测试文本集 A 的特征向量，再使用分类器计算测试文本集 A 的文本分类准确率，即每只青蛙所处位置的适应值；

(4) 按照改进后的 SFLA 算法流程，直到算法迭代次数达到 T 或者满足其他停止条件时，终止算法，并输出最优解 X_g，输出 X_g 各个维度的值为1的特征词，即经过 SFLA 二次优选的特征集合。

基于改进后的 SFLA 的文本特征选择优化算法流程如图 2 所示。

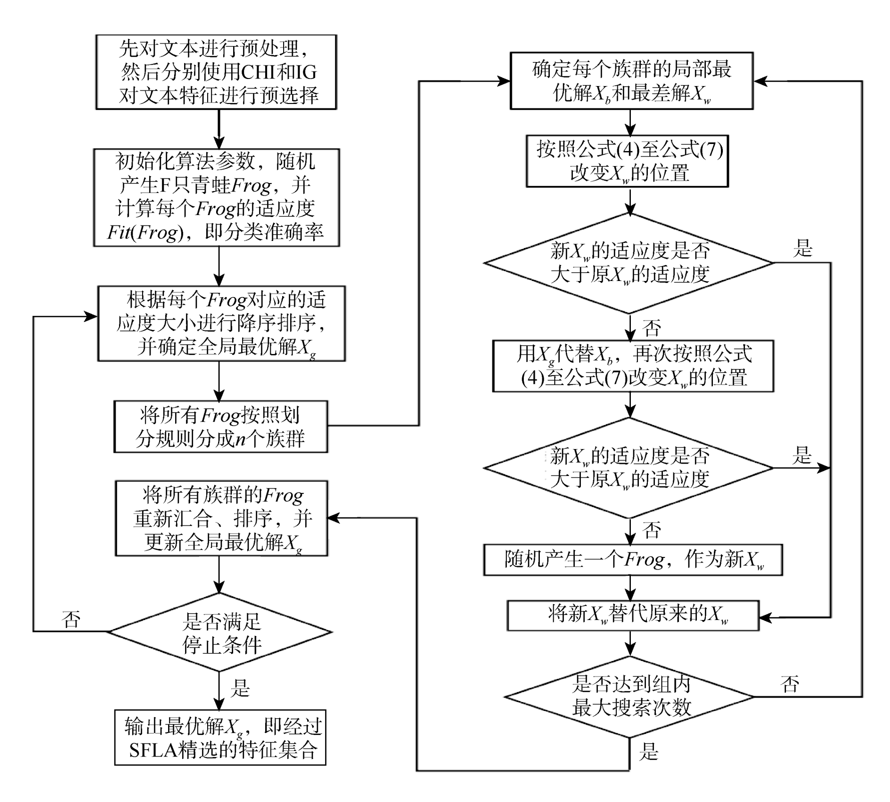


图 2　改进后的 SFLA 的特征选择优化方法流程图

4　实验分析

整个实验主要分为两个部分：第一部分是未使用 SFLA 进行特征优化，即直接将经过传统特征选择方法 CHI 或 IG 选出的特征集合用于文本分类；第二部分则是引入 SFLA 对特征集合进行二次优选，得到高精度的特征集合，并将其用于文本分类，如图 3 所示。

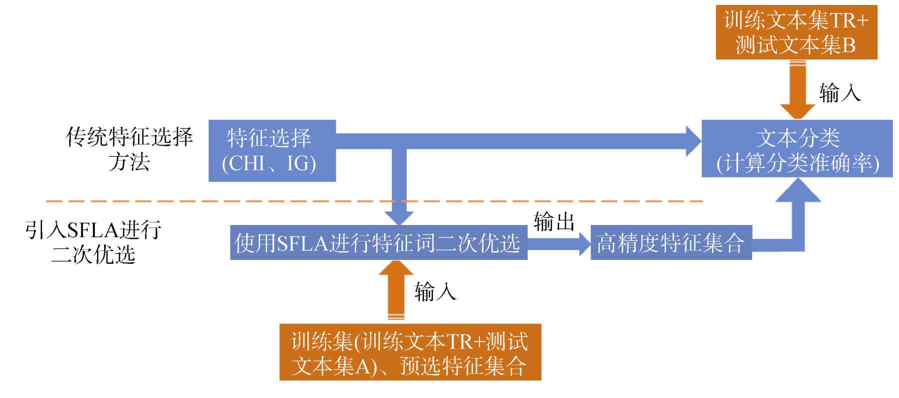


图 3　实验流程图

在直接使用传统特征选择方法 CHI 或 IG 选出的特征集合用于文本分类的过程中，所使用的数据集是训练文本集 TR 和测试文本集 B，用于计算原始特征集合对应的文本分类准确率。

在引入 SFLA 进行特征集合的二次优选过程中，由于需要计算 SFLA 的适应度，即文本分类准确率，所使用的数据集必须包含一个训练文本集和一个测试文本集，因此该过程将训练文本集 TR 和测试文本集 A 作为数据集，即建立模型所需的训练集。而在得到高精度特征集合后，需要计算文本分类准确率，此时则使用测试文本集 B，即评估模型性能的测试集。之所以这两个过程使用到两个测试文本集，是因为：在

使用 SFLA 进行特征优化后，所得到的高精度特征集合可能对测试文本集 A 产生较高的依赖程度，因此无法验证这个高精度特征集合是否对其他测试文本集同样有更好的文本分类效果。因此在使用 SFLA 进行特征优化得到高精度特征集合后，需要使用测试文本集 B 对该特征集合进行评估，检验其分类准确率是否高于原先使用传统特征选择方法得到的准确率。另外，使用 SFLA 进行特征优化的过程中需要多次计算分类准确率，如果在 SFLA 特征优化过程中所使用的测试文本集规模很大，会大大增加时间开销；所以特征优化过程中所采用的测试文本集 A 规模较小，并且测试文本集 A 是从测试文本集 B 的每个类别中各抽取 15% 而组成的数据集。

为了更好地说明算法的有效性，实验分别使用英文和中文数据集。实验一所采用的数据集是路透社语料库 Reuters-21578 的一部分；实验二所采用的数据集是中山大学资讯管理学院智能信息处理实验室语料库的一部分(简称实验室语料库)。

实验所使用的操作系统为 32 位的 Win 10 系统，内存 4GB，i5-2400 处理器，利用 Java 语言编写程序。文本预处理操作使用 Lucene 开源包，分词操作使用中国科学院计算技术研究所分词系统 ICTCLAS[21]。预选特征词分别使用 CHI 和 IG。计算文本特征权重则使用 TF-IDF，使用 SVM 和 KNN 两种分类器进行分类。实验的具体步骤如下：

(1) 将训练文本集 TR 和测试文本集 B 作为数据集，使用 CHI 特征选择方法，分别预选出 100-1200 维(每隔 100 取一个)共 12 个不同维度的特征集合 CHI_100-CHI_1200，并分别计算不同维度下的分类准确率 P_{CHI}；

(2) 使用改进后的二进制 SFLA 对步骤(1)得到的 12 个不同维度的特征集合进行二次优选。优选过程将训练文本集 TR 和测试文本集 A 作为模型的训练集，用于计算每个解的适应度，即分类准确率。最终分别输出 SFLA 的最优解，即 CHI_100-CHI_1200 经过特征词二次优选后的高精度特征集合；

(3) 将步骤(2)得到的二次优选后的高精度特征集合，以训练文本集 TR 和测试文本集 B 作为数据集，分别计算不同维度下的分类准确率 P_{CHI_SFLA}；

(4) 将训练文本集 TR 和测试文本集 B 作为数据集，使用 IG 特征选择方法，分别预选出 100-1200 维共 12 个不同维度的特征集合 IG_100-IG_1200，并分别计算不同维度下的分类准确率 P_{IG}；

(5) 使用改进后的二进制 SFLA 对步骤(4)得到的 12 个不同维度的特征集合进行特征词的二次优选。同步骤(2)，该过程将训练文本集 TR 和测试文本集 A 作为模型的训练集，用于计算每个解的适应度，即分类准确率。最终分别输出 SFLA 的最优解，即 IG_100-IG_1200 各自经过特征词二次优选后的高精度特征集合；

(6) 将步骤(5)得到的二次优选后的高精度特征集合，以训练文本集 TR 和测试文本集 B 作为数据集，分别计算不同维度下的分类准确率 P_{IG_SFLA}；

(7) 在 12 个不同维度下分别比较未使用 SFLA 进行特征词二次优选的准确率 P_{CHI}、P_{IG} 与使用 SFLA 进行特征词二次优选的准确率 P_{CHI_SFLA}、P_{IG_SFLA}，观察使用前后的准确率是否存在较大差别；

(8) 将所有记录的准确率分成两组，分别是：使用 SFLA 前的分类准确率 P_old，使用 SFLA 后的分类准确率 P_new。然后使用配对样本 T 检验，判断两种方法得到结果差异是否存在统计学意义。

4.1 实验一

实验一采用路透社语料库 Reuters-21578，共有 acq、crude、earn、grain、interest、money-fx、ship、trade 这 8 个类别。大测试文本集和训练文本集按 1:2.5 进行划分，各个类别的具体文本数量如表 1 所示。

表 1 Reuters-21578 语料类别分布表

类别	acq	crude	earn	grain	interest	money-fx	ship	trade	总数
训练集	1 596	253	2 840	41	190	206	108	251	5 485
大测试集	696	121	1 083	10	81	87	36	75	2 189

在使用 SVM 分类器时，CHI 或 IG 方法预选出的特征集合经过二次优选后的实验结果如表 2 所示。

将 CHI 和 IG 两组分别绘制成折线图如图 4 和图 5 所示。与表 2 相对应，图 4 和图 5 的横坐标均是指预选特征集合的特征词数量。CHI_SFLA 是指使用 CHI 进行特征词预选，再使用 SFLA 进行二次优选；IG_SFLA 是指使用 IG 进行特征词预选，再使用 SFLA 进行二次优选。在使用 SVM 分类器，Reuters-21578 英文语料库作为数据集时，使用改进后的 SFLA 二次优

选方法明显比传统特征选择方法 CHI 和 IG 能得到更高的分类准确率，并且随着维度的增加，分类准确率的提升幅度有增加的趋势。

表 2　SVM 分类器下 Reuters-21578 各个特征选择方法的分类准确率

特征选择方法 / 维数	CHI (%)	CHI_SFLA (%)	IG (%)	IG_SFLA (%)
100	93.102	92.143	90.132	90.772
200	93.878	92.965	91.366	92.873
300	92.554	92.005	89.082	92.736
400	91.000	94.381	86.249	92.873
500	90.726	94.153	85.381	92.325
600	87.848	92.599	84.651	92.645
700	85.975	93.878	83.919	92.462
800	85.244	93.970	83.645	92.234
900	84.513	93.878	83.326	91.594
1 000	84.011	93.559	82.914	91.640
1 100	83.646	94.107	82.686	93.376
1 200	83.189	94.290	82.412	92.828

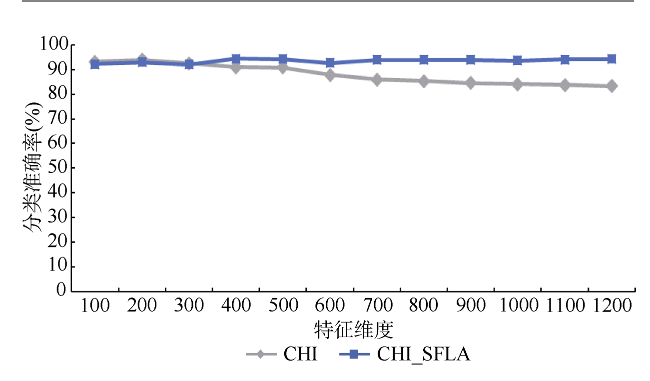


图 4　SVM 分类器下 Reuters-21578 英文语料库的分类准确率(CHI)

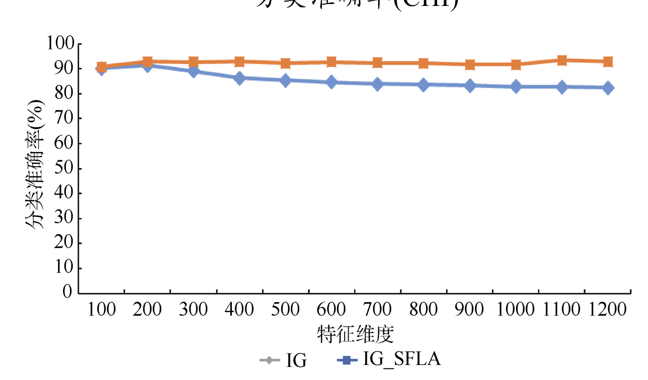


图 5　SVM 分类器下 Reuters-21578 英文语料库的分类准确率(IG)

在使用 KNN 分类器时，CHI 或 IG 方法预选出的特征集合经过二次优选后的实验结果如表 3 所示。

表 3　KNN 分类器下 Reuters-21578 各个特征选择方法的分类准确率

特征选择方法 / 维数	CHI (%)	CHI_SFLA (%)	IG (%)	IG_SFLA (%)
100	90.361	91.914	87.391	90.955
200	88.305	90.909	89.356	90.452
300	87.483	91.275	89.082	90.361
400	86.752	89.630	89.676	89.676
500	87.300	91.366	88.305	88.716
600	87.163	91.594	87.483	89.402
700	86.661	91.138	87.117	89.630
800	85.564	88.671	86.341	89.950
900	84.742	88.031	86.067	89.676
1 000	83.920	88.077	85.062	89.127
1 100	81.361	87.803	84.376	89.493
1 200	81.635	87.163	83.919	89.721

将 CHI 和 IG 两组分别绘制成折线图如图 6 和图 7 所示。

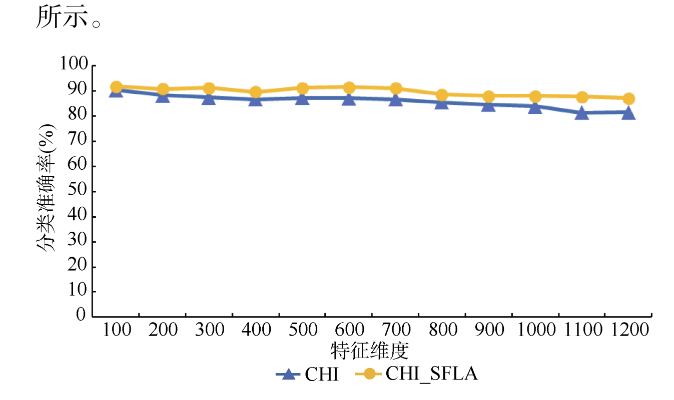


图 6　KNN 分类器下 Reuters-21578 英文语料库的分类准确率(CHI)

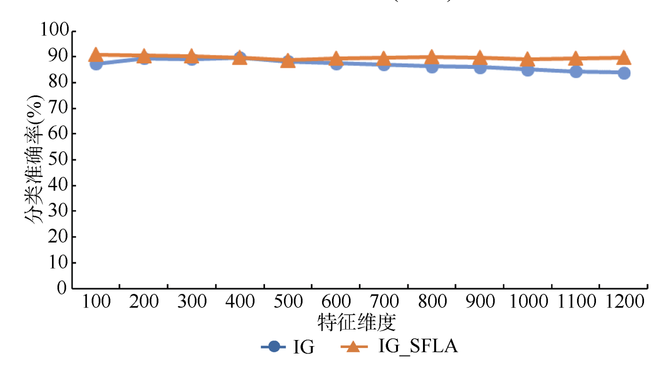


图 7　KNN 分类器下 Reuters-21578 英文语料库的分类准确率(IG)

与表 3 相对应，图 6 和图 7 的横坐标均是指预选特征集合的特征词数量。可以看出，在使用 KNN 分类器，Reuters-21578 英文语料库作为数据集时，在大多数维度中，使用改进后的 SFLA 二次优选方法比传统特征选择方法 CHI 和 IG 能得到较高的分类准确率，但是在 400 维度时，IG_SFLA 所取得的分类准确率跟 IG 的恰好一样，没有提高，但此时经过 IG_SFLA 二次优选后的特征集合的维度小于 400 维度，这也从另一个角度说明通过 IG 预选出来的 400 维度特征集合中存在与分类无关的词项，这一部分词项完全可以剔除。

4.2 实验二

实验室语料库是由中山大学资讯管理学院智能信息处理实验室所收集和整理[22]，共有 13 个类别。本次实验从中选取文本数量较多的 8 个类：education、entertainment、event、finance、game、occultism、sport、technology，从每个类别中随机选取 200 篇文本，共 1 600 篇，作为实验的训练文本集；从剩下的文本集中每个类别随机选取 200 篇文本，共 1 600 篇，作为实验的测试文本集 B，用于对精选后的特征集合的检验；再从 8 个类中每类随机选取 20 篇文本，共 160 篇，作为实验的测试文本集 A。对训练文本集进行文本预处理、分词去重以及去除停用词后，共得到 52 794 个特征词。

在使用 SVM 分类器时，CHI 或 IG 方法预选出的特征集合经过二次优选后的实验结果如表 4 所示。

表 4　SVM 分类器下实验室语料库各个特征选择方法的分类准确率

维数 \ 特征选择方法	CHI (%)	CHI_SFLA (%)	IG (%)	IG_SFLA (%)
100	77.042	77.417	55.667	56.958
200	83.292	85.792	68.667	76.333
300	80.833	86.083	73.833	83.083
400	77.458	84.625	77.083	79.000
500	78.875	85.708	78.708	80.292
600	80.583	86.167	80.083	83.458
700	80.417	86.208	81.167	84.625
800	80.375	85.333	81.833	86.250
900	80.667	85.958	81.417	84.708
1 000	80.750	87.292	81.167	86.667
1 100	80.583	84.667	80.500	82.125
1 200	80.208	86.042	80.250	83.250

将 CHI 和 IG 两组分别绘制成折线图如图 8 和图 9 所示。

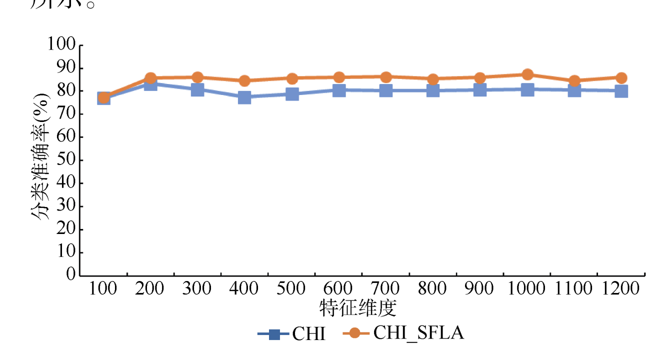


图 8　SVM 分类器下实验室语料库的 CHI_SFLA 和 CHI 的分类准确率

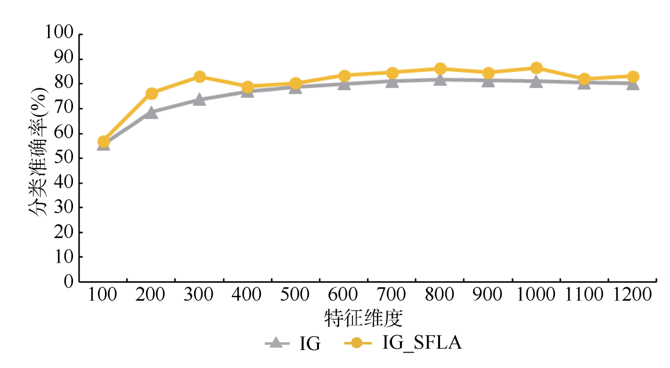


图 9　SVM 分类器下实验室语料库的 IG_SFLA 和 IG 的分类准确率

从图 8 和图 9 可以看出，采用实验室语料库作为数据集时，改进后的 SFLA 二次优选方法比传统特征选择方法 CHI 和 IG 均能得到较高的分类准确率。并且二者都是在维度为 1 000 维的时候取得最高分类准确率。在提高幅度方面，CHI_SFLA 在 400 维度时比 CHI 提高了约 7%，IG_SFLA 在 300 维度时比 IG 提高了约 9%。总体而言，当使用传统特征选择方法所得到的分类准确率较低时，改进后的 SFLA 二次优选方法的优化效果比较明显。

在使用 KNN 分类器时，CHI 或 IG 方法预选出的特征集合经过二次优选后的实验结果如表 5 所示。

将 CHI 和 IG 两组分别绘制成折线图如图 10 和图 11 所示。从图 10 可以看出，在 KNN 分类器下，采用实验室语料库作为数据集时，CHI_SFLA 比 CHI 取得更高的分类准确率，但是在 100 维和 1 000 维时提高幅度不明显，但也达到了降维效果。从图 11 可以看出，在 KNN 分类器下，IG_SFLA 明显比 IG 取得更高的分类准确率，在 1 000 维和 1 100 维时提高幅度达到 12%。

表 5 KNN 分类器下实验室语料库各个特征选择方法的分类准确率

维数 \ 特征选择方法	CHI (%)	CHI_SFLA (%)	IG (%)	IG_SFLA (%)
100	72.125	72.750	52.958	55.583
200	66.750	78.583	65.875	75.125
300	69.250	77.083	65.458	72.917
400	68.458	76.333	67.667	71.917
500	69.083	79.000	67.167	70.917
600	68.167	76.708	65.917	72.292
700	68.083	75.500	64.542	69.917
800	68.750	77.292	60.458	70.458
900	68.167	76.167	57.208	68.833
1 000	70.625	74.708	57.167	69.917
1 100	71.417	77.208	58.667	71.458
1 200	69.958	78.792	60.792	68.750

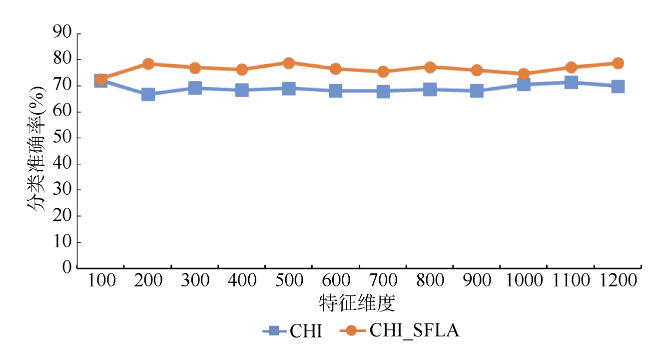


图 10 KNN 分类器下实验室语料库的 CHI_SFLA 和 CHI 的分类准确率

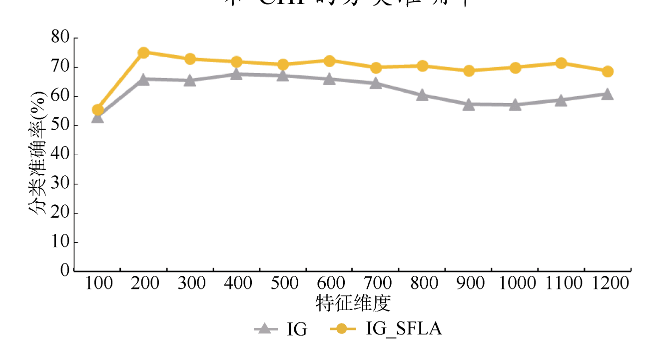


图 11 KNN 分类器下实验室语料库的 IG_SFLA 和 IG 的分类准确率

4.3 配对样本 T 检验

将所有得到的准确率数据分成两组，分别是 P_old 和 P_new，在 SPSS 工具中使用配对样本 T 检验，结果如表 6 所示。

从表 6 可以看出，Sig.=.000<0.01，说明在显著度为 99%的水平下，使用 SFLA 前的分类准确率 P_old 和使用 SFLA 后的分类准确率 P_new 存在显著差异，可见使用 SFLA 进行特征优化选择后，对文本的分类准确率有明显的提升效果。

表 6 配对样本 T 检验结果表格

		成对差分					t	df	Sig. (双侧)
		均值	标准差	均值的标准误	差分的 95%置信区间				
					下限	上限			
对 1	P_old-P_new	–5.39820	3.29716	.33651	–6.06626	–4.73013	–16.042	95	.000

4.4 实验结论

实验一和实验二的结果都说明了基于改进后的 SFLA 的文本特征选择优化算法比传统的 CHI 和 IG 能得到更好的分类效果，说明了改进后的 SFLA 对文本特征二次优选具有较好的可行性和有效性，原因主要是: CHI、IG 等传统的特征选择方法是通过某一数学评价模型从原始特征集合中筛选出具有较好的区分能力和文本代表性的特征集合，即是在统计学角度进行筛选特征集合的，并没有从文本的角度考虑特征词之间的相互影响以及冗余词项对文本分类效果的整体影响。因此使用 CHI 和 IG 所得到的候选特征集合必然存在较多噪声特征项，对分类器的分类效果会造成较大的影响，从而使分类准确率相对较低；而引入改进后的 SFLA 之后，对特征集合进行了二次优选，利用 SFLA 的迭代寻优且收敛性较好的特点，保留具有区分能力的特征词项，并排除部分与分类无关的噪声词项，从而较大程度地提高了文本分类的准确率。

5 结 语

本文从特征选择对文本分类效果的整体影响角度出发，引入了在文本领域应用不多的 SFLA 并尝试将其应用在文本特征选择优化中。通过与传统特征选择

方法 CHI 和 IG 的对比实验可以看到，基于改进后的 SFLA 的文本特征选择优化方法较 CHI 和 IG 能取得更高的分类准确率，主要是因为在算法迭代过程中对预选特征集合去除了较多噪声特征项，降低了噪声特征项对文本分类的影响程度，从而能得到更好的分类效果。然而本文所使用的改进后的 SFLA 相关参数的设置只是基于小规模测试实验得出的结果，下一步将尝试通过对 SFLA 的相关参数进行寻优，找到相关参数的最佳取值范围，使算法结果进一步接近最优解，从而得到更优的高精度特征集合以及更好的分类效果。

参考文献：

[1] 庞观松，蒋盛益．文本自动分类技术研究综述[J]．情报理论与实践，2012，35(2)：123-128．(Pang Guansong, Jiang Shengyi. Text Automatic Classification Technology Research [J]. Information Studies: Theory & Application, 2012, 35(2): 123-128.)

[2] 吴科．基于机器学习的文本分类研究[D]．上海:上海交通大学, 2008. (Wu Ke.A Study on Text Categorization Based on Machine Learning [D]. Shanghai: Shanghai Jiaotong University, 2008.)

[3] 伍建军，康耀红．文本分类中特征选择方法的比较和改进[J]．郑州大学学报：理学版，2007,39(2)：110-113．(Wu Jianjun, Kang Yaohong. Comparison and Improvement of Feature Selection for Text Categorization [J]. Journal of Zhengzhou University: Natural Science Edition, 2007,39(2): 110-113.)

[4] Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]//Proceedings of the 14th International Conference on Machine Learning.San Francisco: Morgan Kaufmann Publishers Inc., 1997: 412-420.

[5] 符发．中文文本分类中特征选择方法的比较[J]．现代计算机：专业版, 2008(6): 43-45. (Fu Fa. Comparison of Feature Selection in Chinese Text Categorization [J]. Modern Computer, 2008(6): 43-45.)

[6] Tabakhi S, Moradi P, Akhlaghian F. An Unsupervised Feature Selection Algorithm Based on Ant Colony Optimization [J]. Engineering Applications of Artificial Intelligence, 2014, 32: 112-123.

[7] 刘亚南．KNN 文本分类中基于遗传算法的特征提取技术研究[D]．北京：中国石油大学, 2011. (Liu Ya'nan. Research of Feature Extraction Technology in KNN Text Classification Based on the Genetic Algorithm [D]. Beijing: China University of Petroleum, 2011.)

[8] 刘逵．基于野草算法的文本特征选择研究[D]．重庆：西南大学，2013．(Liu Kui. An Invasive Weed Optimization Algorithm for Text Feature Selection [D]. Chongqing: Southwest University, 2013.)

[9] Uguz H. A Two-stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm [J]. Knowledge-Based Systems, 2011, 24(7): 1024-1032.

[10] Javed K, Maruf S, Babri H A. A Two-stage Markov Blanket Based Feature Selection Algorithm for Text Classification [J]. Neurocomputing, 2015, 157: 91-104.

[11] Lu Y, Liang M, Ye Z, et al. Improved Particle Swarm Optimization Algorithm and Its Application in Text Feature Selection [J]. Applied Soft Computing, 2015, 35(C): 629-636.

[12] Eusuff M M, Lansey K E. Optimization of Water Distribution Network Design Using the Shuffled Frog Leaping Algorithm [J]. Journal of Water Resources Planning and Management, 2003, 129(3): 210-225.

[13] 崔文华，刘晓冰，王伟，等．混合蛙跳算法研究综述[J]．控制与决策，2012，27(4)：481-486，493．(Cui Wenhua, Liu Xiaobing, Wang Wei, et al. Survey on Shuffled Frog Leaping Algorithm[J]. Control and Decision, 2012, 27(4): 481-486, 493.)

[14] Elbehairy H, Elbeltagi E, Hegazy T, et al. Comparison of Two Evolutionary Algorithms for Optimization of Bridge Deck Repairs [J]. Computer-Aided Civil and Infrastructure Engineering, 2006, 21(8): 561-572.

[15] 陈功贵，李智欢，陈金富，等．含风电场电力系统动态优化潮流的混合蛙跳算法[J]．电力系统自动化，2009，33(4)：25-30．(Chen Gonggui, Li Zhihuan, Chen Jinfu, et al. SFL Algorithm Based Dynamic Optimal Power Flow in Wind Power Integrated System [J]. Automation of Electric Power Systems, 2009, 33(4): 25-30.)

[16] 张沈习，陈楷，龙禹，等．基于混合蛙跳算法的分布式风电源规划[J]．电力系统自动化，2013,37(13)：76-82．(Zhang Shenxi, Chen Kai, Long Yu, et al. Distributed Wind Generator Planning Based Shuffled Frog Leaping Algorithm [J]. Automation of Electric Power Systems, 2013, 37(13): 76-82.)

[17] 余华，黄程韦，金赟，等．基于改进的蛙跳算法的神经网络在语音情感识别中的研究[J]．信号处理，2010，26(9)：1294-1299. (Yu Hua, Huang Chengwei, Jin Yun, et al. Speech Emotion Recognition Based on Modified Shuffled Frog Leaping Algorithm Neural Network [J]. Signal Processing, 2010, 26(9): 1294-1299.)

[18] 许方. 基于混合蛙跳算法的 Web 文本聚类研究[D]. 无锡: 江南大学, 2013. (Xu Fang. Research on Web Text Cluster Algorithm Based on Shuffled Frog-leaping Algorithm [D]. Wuxi: Jiangnan University, 2013.)

[19] 尉建兴, 崔冬华, 宁晓青. 蛙跳算法在 Web 文本聚类技术中的应用[J]. 电脑开发与应用, 2011, 24(5): 35-37. (Yu Jianxing, Cui Donghua, Ning Xiaoqing. Applicatin of Shuffled Frog-leaping Algorithm to Web's Text Cluster Technology [J]. Computer Development & Applications, 2011, 24(5): 35-37.)

[20] Sun X, Wang Z. An Efficient Document Categorization Algorithm Based on LDA and SFL [C]//Proceedings of the 2008 International Seminar on Business and Information Management. IEEE, 2008: 113-115.

[21] NLPIR 汉 语 分 词 系 统 [EB/OL]. [2016-03-17]. http://ictclas.nlpir.org. (NLPIR Chinese Word Segmentation System [EB/OL]. [2016-03-17]. http://ictclas.nlpir.org.)

[22] 路永和, 彭燕虹. 融合实用性与科学性的互联网信息分类体系构建[J]. 图书与情报, 2015(3): 118-124. (Lu Yonghe, Peng Yanhong. The Classification System Construction for Internet Information both Practical and Scientific[J]. Library and Information, 2015(3): 118-124.)

作者贡献声明:

路永和: 提出研究思路和实验建议, 修改论文;
陈景煌: 分析数据, 设计并实现算法程序, 完成实验, 论文撰写以及最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 路永和, 陈景煌. 实验数据集.rar. 从 Reuters-21578 语料库和中山大学资讯管理学院极天智能实验室语料库中选择的一部分文本集作为实验数据集.

[2] 路永和, 陈景煌. 实验输入的预选特征集合.rar. 通过 CHI 和 IG 预选出来的特征词集合.

[3] 路永和, 陈景煌. 实验输出的特征集合.rar. 通过改进的 SFLA 精选出来的特征词集合.

[4] 路永和, 陈景煌. 实验分类结果集.xlsx. 使用 SFLA 精选出来的特征集合后所计算得到的文本分类准确率.

收稿日期: 2016-09-30
收修改稿日期: 2016-12-12

Optimizing Feature Selection Method for Text Classification with Shuffled Frog Leaping Algorithm

Lu Yonghe Chen Jinghuang
(School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract: [**Objective**] This paper introduces the shuffled frog leaping algorithm (SFLA) to remove the irrelevant terms from the texts, and optimizes the feature selection method to improve the accuracy of text classification. [**Methods**] First, we used CHI and IG techniques to pre-select different dimensions of feature terms, and then adopted the modified SFLA to refine the text features' list. Second, we used a frog to represent a feature selection rule, and applied the classification precision as the fitness function. Finally, the SVM and KNN classifier were adopted to calculate the classification precision. [**Results**] The modified SFLA had better performance in classification precision than CHI and IG, and the highest increasing rate was 12%. [**Limitations**] The feature over fitting occured in small portion of space dimensions. [**Conclusions**] Using feature preselection and the modified SFLA could effectively exclude irrelevant or invalid terms, and then improve the precision of feature selection.

Keywords: Feature Selection Text Classification Shuffled Frog Leaping Algorithm